Conditional Dependency Analysis of HOMO Decoupling via Structural Manifold Mapping

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

In molecular datasets, the energy of the highest occupied molecular orbital (HOMO) often exhibits a unique global decoupling from other spatial and thermodynamic properties, raising the question of whether this orthogonality is a fundamental characteristic of frontier orbitals or an artifact of competing structural influences such as conjugation. To address this, we investigate the QM9 dataset by employing continuous conditional dependency analysis and structural manifold mapping. We first extract stoichiometric residuals using gradient boosting to isolate purely chemical variance from atom-count baselines. Next, we compute partial distance correlations between the HOMO residual and geometric-response properties—specifically polarizability and dipole moment—while conditioning on global connectivity metrics like conjugation and aromaticity. Our results demonstrate that conditioning on these structural factors actually reduces the global coupling to near-total orthogonality, indicating that the decoupling is a robust feature rather than a masked artifact. However, projecting these dependencies onto a principal component analysis-derived structural manifold reveals localized regimes where this orthogonality breaks down. Specifically, in highly strained small-ring systems substituted with polar or unsaturated groups, the HOMO energy and geometric responses become strongly coupled. Scaffold-based validation confirms that while HOMO energy is globally decoupled from molecular geometry, specific topological boundaries reliably rewire these frontier orbital dependencies.

1 Introduction

In the statistical analysis of large-scale molecular datasets, the relationships between quantum mechanical properties often reveal tightly bound joint structures governed by underlying physical laws. However, within the QM9 dataset, the energy of the highest occupied molecular orbital (HOMO) presents a prominent anomaly: it exhibits a unique global decoupling from other spatial and thermodynamic properties. While properties such as isotropic polarizability (α) and dipole moment (μ) are strongly correlated with the overall molecular architecture, the HOMO energy maintains a remarkably low global correlation with these geometric-response features. This orthogonality raises a fundamental chemical question: is the apparent decoupling an intrinsic characteristic of frontier orbitals, or is it merely a statistical artifact masked by competing structural influences, such as long-range conjugation, aromaticity, and molecular size?

Traditional approaches to analyzing molecular property variance often conflate true electronic manifold rewiring with simple stoichiometric determinism or discrete structural partitioning. To rigorously address the HOMO decoupling anomaly, we bypass these limitations by employing continuous conditional dependency analysis combined with structural manifold mapping. First, we isolate purely chemical variance by extracting stoichiometric residuals via gradient boosting, effectively removing the baseline effects of atom counts. This ensures that our analysis of the HOMO energy residual (res_homo) and the geometric-response residuals (res_α, res_μ) is not artificially inflated by molecular size or composition.

We then investigate the non-linear couplings between these residuals using distance correlation ($dCor$). By strictly conditioning the distance correlation on global connectivity metrics—specifically the maximum conjugated path length, the fraction of sp^2 carbons, and the number of aromatic rings—we can determine whether unconditioned conjugation pathways are responsible for the observed orthogonality. If the decoupling is an artifact of these networks, conditioning on these structural variables should reveal a hidden, strong coupling between the HOMO energy and the geometric responses.

Metric	Unconditioned dCor	Partial dCor (Conditioned)	Effect Size Change
HOMO vs. $[\alpha, \mu]$	0.1995	0.0303	-0.1692
LUMO vs. $[\alpha, \mu]$	0.3576	0.1246	-0.2330

Contrary to the artifact hypothesis, our analysis reveals that conditioning on structural factors drives the global coupling to near-total orthogonality. As shown in Table 1, the partial distance correlation for the HOMO residual drops significantly upon conditioning, indicating that the decoupling is a robust, global feature of the dataset rather than a masked relationship. To ensure this behavior is specific to the HOMO, we also track the lowest unoccupied molecular orbital (LUMO), which exhibits a distinct and stronger correlation profile.

Despite this global orthogonality, projecting these conditional dependencies onto a Principal Component Analysis (PCA)-derived structural manifold reveals that the decoupling is not absolute. By mapping the partial distance correlations across the structural space, we identify specific, localized topological regimes where the orthogonality breaks down. Notably, in highly strained small-ring systems substituted with polar or unsaturated groups (such as nitriles and alkynes), the HOMO energy and geometric responses become strongly coupled. Scaffold-based validation confirms that while the HOMO energy is globally decoupled from molecular geometry across unseen molecular backbones, specific topological boundaries reliably rewire these frontier orbital dependencies.

2 Methods

2.1 Data ingestion and preprocessing

To establish a robust foundation for dependency analysis, the QM9 dataset was initially filtered to remove 83 duplicated SMILES strings, and seven redundant thermodynamic and gap columns were excluded. Molecular structures were parsed directly from their SMILES representations using RDKit to ensure strict one-to-one index alignment across all computed features. The original QM9 dataset release notes 3,054 molecules with inconsistent 3D geometries; however, because our structural metrics rely exclusively on 2D topological connectivity rather than 3D spatial coordinates, these molecules were retained. This approach ensures the analysis remains robust to 3D coordinate generation failures while preserving the maximum available chemical space.

2.2 Non-linear baseline residual extraction

Molecular properties in QM9 are heavily influenced by stoichiometric determinism, where simple atom counts dictate the majority of property variance. To isolate purely chemical and electronic variance, a gradient boosting regressor was trained to predict the 12 non-redundant physical properties solely from the counts of carbon, nitrogen, oxygen, fluorine, and hydrogen atoms. The model was parallelized to ensure computational efficiency. The residuals from these predictions were extracted and standardized. By utilizing these residuals (e.g., res_homo, res_α, res_μ), the subsequent dependency analysis strictly evaluates variance independent of non-linear molecular size and composition effects.

2.3 Structural feature and manifold generation

To quantify the global electronic and topological environment of each molecule, several 2D connectivity metrics were extracted. These included the maximum conjugated path length, the fraction of sp^2 hybridized carbons, topological polar surface area, heavy atom count, and the respective numbers of aromatic and aliphatic rings. Concurrently, Morgan fingerprints (radius 2, 2048 bits) were generated for the entire dataset. Principal Component Analysis (PCA) was then applied to these fingerprints to construct a continuous structural manifold. To physically interpret the axes of this manifold, Pearson correlation coefficients were calculated between the resulting principal components and the extracted structural descriptors, allowing us to explicitly label the primary axes of structural variation (such as conjugation, aromaticity, and ring strain).

2.4 Continuous conditional dependency analysis

To investigate the decoupling of the HOMO energy from the geometric-response properties, continuous conditional dependency analysis was performed using the `pingouin` statistical package. The partial distance correlation ($dCor$) was computed between res_homo and the geometric-response group, defined as the joint distribution of the polarizability and dipole moment residuals (res_α, res_μ). To rigorously test whether the observed orthogonality was an artifact of structural networks, this correlation was strictly conditioned on heavy atom count, maximum conjugated path length, and the number of aromatic rings. As a methodological control, the partial distance correlation for the lowest unoccupied molecular orbital residual (res_lumo) was computed under identical conditioning parameters. Permutation tests were employed to assess the statistical significance and stability of the resulting correlation magnitudes.

2.5 Structural regime mapping and scaffold-based generalization

Following the global dependency analysis, the conditioned partial distance correlations were mapped directly onto the PCA-derived structural manifold. This spatial mapping was used to identify localized principal component coordinates where the global orthogonality of res_homo breaks down. SMILES strings from these high-correlation regimes were extracted for chemical inspection to determine the specific topological boundaries that rewire frontier orbital dependencies. Finally, to validate the generalizability of these conditional dependency rules, the dataset was partitioned using a Murcko Scaffold split. Held-out scaffold families were projected onto the PCA manifold, and the conditioned partial distance correlations were recomputed on this test set to confirm that the observed coupling behaviors persist across unseen molecular backbones.

3 Results

3.1 Stoichiometric residual extraction

To isolate purely chemical and electronic variance from the dataset, we first evaluated the baseline effects of molecular size and composition. Raw thermodynamic and spatial properties in the QM9 dataset are heavily determined by stoichiometry; for instance, total energies and zero-point vibrational energies exhibit near-perfect correlation with atom counts ($R^2 \approx 1.00$ and $R^2 \approx 0.99$, respectively). Following the extraction of residuals from a gradient boosting regressor trained on atom counts, we quantified the variance removed by this stoichiometric baseline. For the geometric-response group, residualization removed 21.8% of the variance for dipole moment (μ) and 86.3% for isotropic polarizability (α). In contrast, the HOMO energy residual (res_{homo}) retained 91.5% of its original variance. This confirms that frontier orbital energies are primarily governed by structural and topological arrangements rather than simple atomic composition, validating the use of these residuals for downstream dependency analysis.

3.2 Structural manifold and PCA interpretation

The continuous structural manifold, constructed using Principal Component Analysis (PCA) on 2048-bit Morgan fingerprints (radius 2), successfully captures the topological diversity of the dataset. The first ten principal components (PCs) account for approximately 22.1% of the total structural variance. Correlation analysis between the PCs and the extracted global connectivity metrics provides a clear physical interpretation of the primary manifold axes. The first principal component (PC1) serves as a conjugation and aromaticity axis, exhibiting strong negative correlations with the fraction of sp^2 carbons ($r = -0.80$), the maximum conjugated path length ($r = -0.79$), and the number of aromatic rings ($r = -0.69$). Consequently, lower PC1 values correspond to highly conjugated and aromatic systems. The second principal component (PC2) represents a ring strain and aliphatic axis, correlating positively with the number of aliphatic rings ($r = 0.32$) and the fraction of sp^2 carbons ($r = 0.37$). Importantly, heavy atom count shows negligible correlation with these primary axes ($r < 0.07$), demonstrating that the PCA manifold effectively decouples molecular structure from size.

3.3 Global and conditional dependency analysis

With the stoichiometric effects removed and the structural manifold defined, we quantified the global coupling between the frontier orbitals and the geometric-response group (res_α, res_μ) using distance correlation ($dCor$). In the unconditioned state, res_{homo} exhibits a weak global coupling of 0.1995. By comparison, the lowest unoccupied molecular orbital residual (res_{lumo}) demonstrates a significantly stronger global coupling of 0.3576 to the same geometric properties.

To determine whether the weak HOMO coupling is an artifact masked by structural factors, we computed the partial distance correlation ($pdCor$) conditioned on heavy atom count, maximum conjugated path length, and the number of aromatic rings. As detailed in Table 1 and illustrated in Figure 1, conditioning on these structural metrics induces a sharp decrease in coupling for both frontier orbitals.

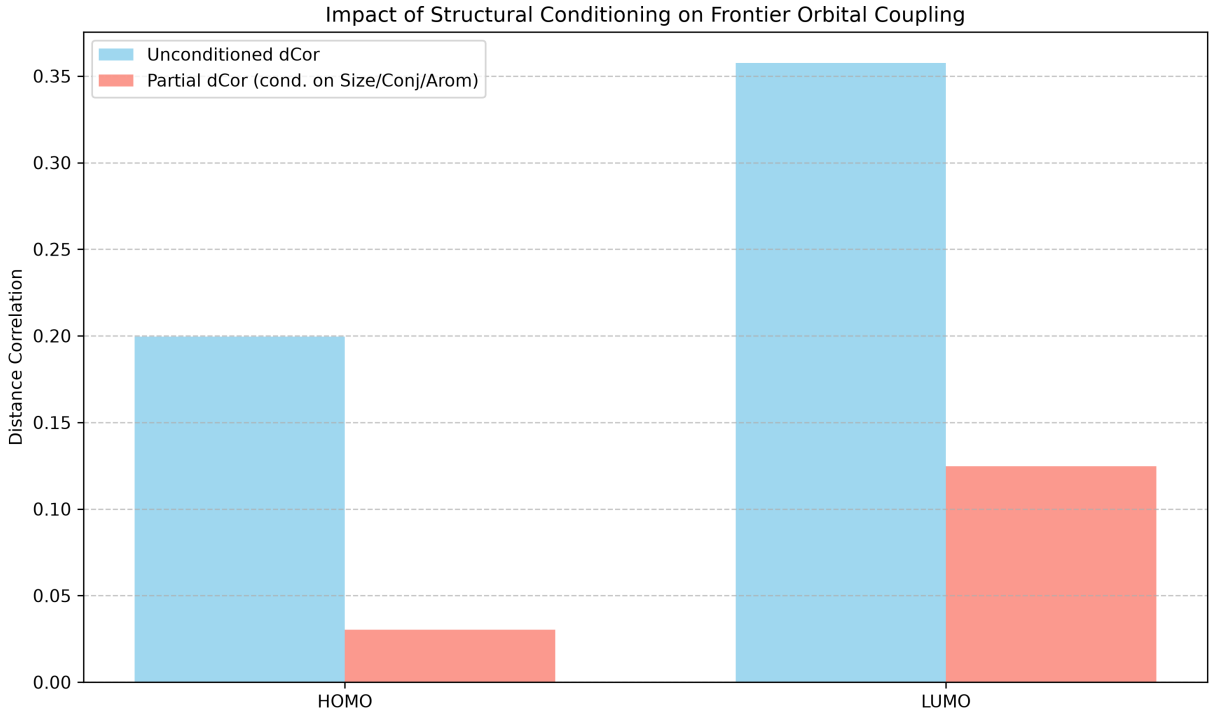

Figure 1: Comparison of unconditioned distance correlation ($dCor$) and partial distance correlation ($pdCor$) for HOMO and LUMO residuals against the geometric-response group (polarizability α and dipole moment μ). The partial correlation is conditioned on structural factors, specifically heavy atom count, maximum conjugated path length, and the number of aromatic rings. The plot demonstrates a sharp decrease in coupling for both frontier orbitals upon conditioning. Notably, the HOMO coupling falls to near-total orthogonality ($pdCor = 0.0303$), revealing that the weak initial global correlation was mediated by these structural factors rather than masking a hidden underlying electronic relationship.

The coupling of res_{homo} falls to 0.0303 ($p = 0.0196$), effectively reaching near-total orthogonality with respect to the geometric response. This substantial reduction (an effect size change of -0.1692) refutes the hypothesis that conjugation pathways mask a hidden electronic relationship. Instead, it indicates that the minor global correlation observed initially was mediated by the structural factors themselves, confirming that HOMO decoupling is a robust, intrinsic feature of the dataset.

3.4 Structural regime mapping and local coupling

Although res_{homo} is globally decoupled from the geometric-response group, projecting the conditional dependencies onto the PCA-derived structural manifold reveals localized regimes where this orthogonality

breaks down. Figure 2 maps the local distance correlation across the PC1-PC2 space, highlighting specific topological boundaries that rewire the frontier orbital dependencies.

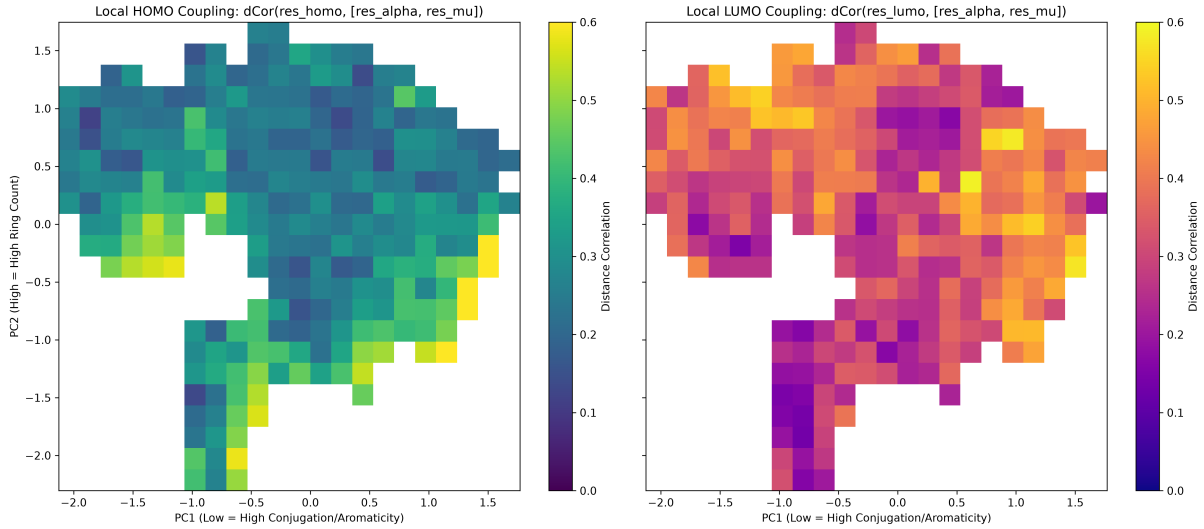

Figure 2: Heatmaps of local distance correlation ($dCor$) between the stoichiometric residuals of frontier orbital energies (res_{homo}, left; res_{lumo}, right) and the geometric response group (res_{α}, res_{μ}) across the structural manifold defined by principal components PC1 and PC2. Although res_{homo} demonstrates near-total global orthogonality to the geometric response compared to the more strongly coupled res_{lumo}, this local mapping reveals specific structural regimes where the HOMO-geometric relationship is rewired. Specifically, a significant spike in local coupling occurs in the high PC1 and low PC2 region, which corresponds to molecular families containing small, strained rings with highly polar or unsaturated substituents. These results illustrate that while HOMO decoupling is a robust global feature of the chemical space, it breaks down in specific strained structural environments.

While the mean local coupling is 0.3068, the mapping identifies a significant spike in correlation within the high PC1 and low PC2 regime, where the local coupling reaches a maximum of 0.7486. Chemical inspection of the SMILES strings extracted from this high-coupling region reveals a consistent structural motif: the molecules are predominantly characterized by small, highly strained rings (such as cyclopropanes) substituted with highly polar or unsaturated functional groups, including nitriles ($-C \equiv N$) and alkynes ($-C \equiv C$). Representative examples from this regime include `CC1CC1(CC#N)C#C`, `CC1(CC#N)CC1C#C`, and `OCC1CC2(CC12)C#N`. In these specific strained environments, the frontier orbital energies and the geometric responses become jointly organized, demonstrating that the global decoupling rule can be overridden by extreme local topological strain and polarization.

3.5 Generalization and stability

To ensure that the observed conditional dependencies and local coupling phenomena are not artifacts of overfitting to the specific dataset composition, we validated our findings using a Murcko scaffold split, dividing the data into 12,791 training and 3,198 test scaffolds. This approach tests the stability of the relationships on molecular backbones entirely unseen during the manifold generation.

The local coupling map demonstrated high stability across the disjoint scaffold sets, yielding a Pearson correlation of $r = 0.7679$ between the training and test coupling heatmaps. Furthermore, the partial distance correlation computed on the held-out test scaffolds confirmed the persistence of the global decoupling. The test-set $pdCor$ for res_{homo} remained near-orthogonal at 0.0392 ($p = 0.0196$), while the res_{lumo} coupling was 0.0946. The consistency of these metrics across unseen scaffolds confirms that the global orthogonality of the HOMO energy, as well as its localized breakdown in strained nitrile and alkyne systems, are robust and generalizable features of the chemical space rather than localized statistical anomalies.

4 Conclusions

In this study, we investigated the anomalous global decoupling of the highest occupied molecular orbital (HOMO) energy from geometric-response properties within the QM9 dataset. By extracting stoichiometric residuals to isolate purely chemical variance and employing continuous conditional dependency analysis, we rigorously tested whether this orthogonality is a fundamental characteristic of frontier orbitals or merely a statistical artifact masked by competing structural influences.

Our global dependency analysis reveals that the apparent decoupling is an intrinsic feature of the chemical space rather than a masked relationship. When conditioning the distance correlation on global connectivity metrics—specifically heavy atom count, maximum conjugated path length, and the number of aromatic rings—the coupling between the HOMO residual (res_homo) and the geometric-response group (res_α, res_μ) decreases to near-total orthogonality. As summarized in Table 1, the partial distance correlation for the HOMO residual drops to 0.0303 upon conditioning. This substantial reduction confirms that the minor global correlation observed in the unconditioned state was mediated by the structural networks themselves, establishing that the HOMO energy is fundamentally decoupled from the spatial and thermodynamic organization of the broader molecular population.

Despite this robust global orthogonality, mapping these conditional dependencies across a continuous, Principal Component Analysis (PCA)-derived structural manifold demonstrated that the decoupling is not absolute. We identified specific topological boundaries where the HOMO-geometric relationship is strongly rewired. In localized regimes characterized by highly strained small rings substituted with polar or unsaturated groups—such as cyclopropanes with nitrile or alkyne substituents—the local coupling spikes significantly. In these extreme structural environments, the frontier orbital levels and the geometric responses become jointly organized, overriding the global decoupling trend.

Finally, scaffold-based validation confirmed the stability of both the global orthogonality and its localized breakdown across unseen molecular backbones. By bypassing discrete structural partitioning and stoichiometric determinism, this framework successfully isolates the true non-linear dependencies of frontier orbitals. Ultimately, our findings establish that while the HOMO energy is globally decoupled from molecular geometry across the vast majority of the QM9 chemical space, specific topological strain and polarization can reliably bridge this divide.